

Express Mail No. EL636048409US

PATENT APPLICATION OF

LI DENG, XUEDONG HUANG, AND MICHAEL D.  
PLUMPE

ENTITLED

PATTERN RECOGNITION TRAINING METHOD AND  
APPARATUS USING INSERTED NOISE FOLLOWED BY  
NOISE REDUCTION

Docket No. M61.12-0315

005707 05688960

## BACKGROUND OF THE INVENTION

A pattern recognition system, such as a speech recognition system, takes an input signal and attempts to decode the signal to find a pattern represented by the signal. For example, in a speech recognition system, a speech signal (often referred to as a test signal) is received by the recognition system and is decoded to identify a string of words represented by the speech signal.

Before a model can be used to decode an incoming signal, it must be trained. This is typically done by measuring input training signals generated from a known training pattern. For example, in speech recognition, a collection of speech signals is generated by speakers reading from a known text. These speech signals are then used to train the models.

5

10

20

25

reduction techniques are then applied to the testing data to bring the testing data closer to the clean condition of the training data. However, current noise reduction techniques are imperfect and cannot  
5 remove all of the noise in the test data. Because of this, the training data and the testing data remain mismatched even after the noise reduction.

Thus, the prior art techniques for matching training data noise to testing data noise are less  
10 than ideal.

#### SUMMARY OF THE INVENTION

A method and apparatus for training and using a pattern recognition model are provided. Under the invention, additive noise that matches  
15 noise expected in a test signal is included in a training signal. The noisy training signal is passed through one or more noise reduction techniques to produce pseudo-clean training data. The pseudo-clean training data is then used to train the pattern  
20 recognition model.

When the test signal is received, it is passed through the same noise reduction techniques used on the noisy training signal. This produces pseudo-clean test data that better matches the  
25 pseudo-clean training data. The pseudo-clean test data is then applied to the pattern recognition model to determine the most likely pattern represented by the test signal.

Under one embodiment, sets of training data  
30 are produced with each set containing a different

009999-101000

type of noise. For example, one set may include fan noise from a computer while another set may include keyboard noise. Under such embodiments, each set of training data may be passed through the same noise reduction techniques or different sets of training data may be passed through different noise reduction techniques.

Under one embodiment, when different noise reduction techniques are used for different sets of training data, the noise in the test data is sampled to identify a particular set of training data that contains a similar type of noise. The noise reduction technique applied to the best matching training data is then applied to the test data to form the pseudo-clean test data.

In other embodiments where different noise reduction techniques are used for different sets of training data or for the same set of training data, the test data is passed through the different noise reduction techniques producing multiple different versions of pseudo-clean test data. Each of these separate forms of pseudo-clean test data is then applied to the models to determine a probability for a pattern. The proper noise reduction technique to apply to the test data is then implicitly selected by selecting the form or combination of forms of the pseudo-clean test data that produces the highest probability pattern.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of one computing environment in which the present invention may be practiced.

5           FIG. 2 is a block diagram of an alternative computing environment in which the present invention may be practiced.

          FIG. 3 is a flow diagram of one embodiment of a method of pattern recognition under the present  
10          invention.

          FIG. 4 is a block diagram of a pattern recognition system under one embodiment of the present invention.

          FIG. 5 is a graph of a speech signal in the  
15          time domain.

          FIG. 6 is the frequency spectrum of noise in a speech signal.

          FIG. 7 is a block diagram of a noise reduction technique used in one embodiment of the  
20          present invention.

          FIG. 8 is a flow diagram for training sets of training data containing different types of noise under one embodiment of the present invention.

          FIG. 9 is a graph of model probability  
25          distributions for different sets of training.

          FIG. 10 is a graph of a combined model probability for the probabilities of FIG. 9.

          FIG. 11 is a graph of the model probability distributions of FIG. 9 after the application of

009307-10160

noise reduction techniques of one embodiment of the present invention.

FIG. 12 is a graph of a combined model probability for the probabilities of FIG. 11.

5           FIG. 13 provides a block diagram of a noise reduction module under an embodiment of the invention using sets of training data.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

FIG. 1 illustrates an example of a suitable  
10   computing system environment 100 on which the invention may be implemented. The computing system environment 100 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or  
15   functionality of the invention. Neither should the computing environment 100 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the exemplary operating environment 100.

20           The invention is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well known computing systems, environments, and/or configurations that may be suitable for use with the  
25   invention include, but are not limited to, personal computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe  
30   computers, distributed computing environments that

005534 101600

The invention may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote computer storage media including memory storage devices.

With reference to FIG. 1, an exemplary system for implementing the invention includes a general-purpose computing device in the form of a computer 110. Components of computer 110 may include, but are not limited to, a processing unit 120, a system memory 130, and a system bus 121 that couples various system components including the system memory to the processing unit 120. The system bus 121 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel



Computer 110 typically includes a variety of computer readable media. Computer readable media can be any available media that can be accessed by computer 110 and includes both volatile and nonvolatile media, removable and non-removable media. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer 100. Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal"

means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes  
5 wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, FR, infrared and other wireless media. Combinations of any of the above should also be included within the scope of computer readable media.

10           The system memory 130 includes computer storage media in the form of volatile and/or nonvolatile memory such as read only memory (ROM) 131 and random access memory (RAM) 132. A basic input/output system 133 (BIOS), containing the basic  
15 routines that help to transfer information between elements within computer 110, such as during start-up, is typically stored in ROM 131. RAM 132 typically contains data and/or program modules that are immediately accessible to and/or presently being  
20 operated on by processing unit 120. By way of example, and not limitation, FIG. 1 illustrates operating system 134, application programs 135, other program modules 136, and program data 137.

25           The computer 110 may also include other removable/non-removable volatile/nonvolatile computer storage media. By way of example only, FIG. 1 illustrates a hard disk drive 141 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 151 that reads from or writes  
30 to a removable, nonvolatile magnetic disk 152, and an

003701-03390

5  
10  
15  
20  
25

The computer 110 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be a personal computer, a hand-held device, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 110. The logical connections depicted in FIG. 1 include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices,

The computer 110 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be a personal computer, a hand-held device, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 110. The logical connections depicted in FIG. 1 include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices,

enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the computer 110 is connected to the LAN 171 through a network interface or adapter 170. When used in a WAN networking environment, the computer 110 typically includes a modem 172 or other means for establishing communications over the WAN 173, such as the Internet. The modem 172, which may be internal or external, may be connected to the system bus 121 via the user input interface 160, or other appropriate mechanism. In a networked environment, program modules depicted relative to the computer 110, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, FIG. 1 illustrates remote application programs 185 as residing on remote computer 180. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

FIG. 2 is a block diagram of a mobile device 200, which is an exemplary computing environment. Mobile device 200 includes a microprocessor 202, memory 204, input/output (I/O) components 206, and a communication interface 208 for communicating with remote computers or other mobile devices. In one embodiment, the afore-mentioned components are coupled for communication with one another over a suitable bus 210.

Memory 204 includes an operating system 212, application programs 214 as well as an object store 216. During operation, operating system 212 is preferably executed by processor 202 from memory 204. Operating system 212, in one preferred embodiment, is a WINDOWS® CE brand operating system commercially available from Microsoft Corporation. Operating system 212 is preferably designed for mobile devices, and implements database features that can be utilized by applications 214 through a set of exposed application programming interfaces and methods. The objects in object store 216 are maintained by applications 214 and operating system 212, at least partially in response to calls to the exposed application programming interfaces and methods.

Communication interface 208 represents numerous devices and technologies that allow mobile device 200 to send and receive information. The devices include wired and wireless modems, satellite receivers and broadcast tuners to name a few. Mobile device 200 can also be directly connected to a

computer to exchange data therewith. In such cases, communication interface 208 can be an infrared transceiver or a serial or parallel communication connection, all of which are capable of transmitting  
5 streaming information.

Input/output components 206 include a variety of input devices such as a touch-sensitive screen, buttons, rollers, and a microphone as well as a variety of output devices including an audio  
10 generator, a vibrating device, and a display. The devices listed above are by way of example and need not all be present on mobile device 200. In addition, other input/output devices may be attached to or found with mobile device 200 within the scope  
15 of the present invention.

Under the present invention, an apparatus and method are provided that improve the matching of noise between training data and test data. FIG. 3 shows one embodiment of a method for performing such  
20 matching.

In step 300 of FIG. 3, raw training data is created that includes anticipated additive noise. This anticipated additive noise is similar to the noise that is expected to be present in the test  
25 data. This anticipated additive noise can be placed in the training data by having a trainer speak in a noisy environment such as a train platform, a car, or an industrial environment. In other embodiments, the trainer speaks in a relatively noiseless environment  
30 and additive noise is then added to the "clean"

009707-0565950

5

10

15

25

30



5 By applying the same noise reduction techniques to the test data that were applied to the training data, the present invention brings the test data into a pseudo-clean state that more closely matches the pseudo-clean state of the training data.

FIG. 4 provides a block diagram of hardware components and program modules found in the general computing environments of FIGS. 1 and 2 that are particularly relevant to an embodiment of the present invention used for speech recognition. In FIG. 4, an input speech signal from a speaker 400 and additive noise 402 are converted into an electrical signal by a microphone 404, which is connected to an analog-to-digital (A-to-D) converter 406. Although additive noise 402 is shown entering through microphone 404 in the embodiment of FIG. 4, in other embodiments, additive noise 402 may be added to the input speech signal as a digital signal after A-to-D converter 406.

A-to-D converter 406 converts the analog signal from microphone 404 into a series of digital values. In several embodiments, A-to-D converter 406

samples the analog signal at 16 kHz and 16 bits per sample, thereby creating 32 kilobytes of speech data per second.

The digital data created by A-to-D  
5 converter 406 is provided to a noise reduction module  
408, which removes some of the noise in the digital  
signal using one or more noise reduction techniques.  
Such noise reduction techniques include but are not  
limited to Spectral Subtraction or Stereo Piecewise  
10 Linear Compensation for Environments (SPLICE).

The output of noise reduction module 408 is  
provided to feature extractor 400, which extracts a  
feature from the digital speech signal. Examples of  
feature extraction modules include modules for  
15 performing Linear Predictive Coding (LPC), LPC  
derived cepstrum, Perceptive Linear Prediction (PLP),  
Auditory model feature extraction, and Mel-Frequency  
Cepstrum Coefficients (MFCC) feature extraction.  
Note that the invention is not limited to these  
20 feature extraction modules and that other modules may  
be used within the context of the present invention.

The feature extraction module receives the  
stream of digital values from noise reduction module  
408 and produces a stream of feature vectors that are  
25 each associated with a frame of the speech signal.  
In many embodiments, the centers of the frames are  
separated by 10 milliseconds.

Note that although noise reduction module  
408 is shown before feature extractor 400 in the  
30 embodiment of FIG. 4, in other embodiments, noise

0055050-404600

The stream of feature vectors produced by the extraction module is provided to a decoder 412, which identifies a most likely sequence of words based on the stream of feature vectors, a lexicon 414, a language model 416, and an acoustic model 418.

In some embodiments, acoustic model 418 is a Hidden Markov Model consisting of a set of hidden states. Each linguistic unit represented by the model consists of a subset of these states. For example, in one embodiment, each phoneme is constructed of three interconnected states. Each state has an associated set of probability distributions that in combination allow efficient computation of the likelihoods against any arbitrary sequence of input feature vectors for each sequence of linguistic units (such as words). The model also includes probabilities for transitioning between two neighboring model states as well as allowed transitions between states for particular linguistic units. By selecting the states that provide the highest combination of matching probabilities and transition probabilities for the input feature vectors, the model is able to assign linguistic units to the speech. For example, if a phoneme was constructed of states 0, 1 and 2 and if the first three frames of speech matched state 0, the next two matched state 1 and the next three matched state 2,

the model would assign the phoneme to these eight frames of speech.

Note that the size of the linguistic units can be different for different embodiments of the present invention. For example, the linguistic units may be senones, phonemes, noise phones, diphones, triphones, or other possibilities.

In other embodiments, acoustic model 418 is a segment model that indicates how likely it is that a sequence of feature vectors would be produced by a segment of a particular duration. The segment model differs from the frame-based model because it uses multiple feature vectors at the same time to make a determination about the likelihood of a particular segment. Because of this, it provides a better model of large-scale transitions in the speech signal. In addition, the segment model looks at multiple durations for each segment and determines a separate probability for each duration. As such, it provides a more accurate model for segments that have longer durations. Several types of segment models may be used with the present invention including probabilistic-trajectory segmental Hidden Markov Models.

Language model 416 provides a set of likelihoods that a particular sequence of words will appear in the language of interest. In many embodiments, the language model is based on a text database such as the North American Business News (NAB), which is described in greater detail in a

publication entitled CSR-III Text Language Model, University of Penn., 1994. The language model may be a context-free grammar or a statistical N-gram model such as a trigram. In one embodiment, the language  
5 model is a compact trigram model that determines the probability of a sequence of words based on the combined probabilities of three-word segments of the sequence.

Based on the acoustic model, the language  
10 model, and the lexicon, decoder 412 identifies a most likely sequence of words from all possible word sequences. The particular method used for decoding is not important to the present invention and any of several known methods for decoding may be used.

15 The most probable sequence of hypothesis words is provided to a confidence measure module 420. Confidence measure module 420 identifies which words are most likely to have been improperly identified by the speech recognizer, based in part on a secondary  
20 frame-based acoustic model. Confidence measure module 420 then provides the sequence of hypothesis words to an output module 422 along with identifiers indicating which words may have been improperly identified. Those skilled in the art will recognize  
25 that confidence measure module 420 is not necessary for the practice of the present invention.

Acoustic model 418 above is trained by a  
trainer 424 based on a training text 426 and the  
features extracted by feature extractor 410 from one  
30 or more training speech signals associated with

0066950 10160  
00507 0566950

training text 426. Any suitable training method that is appropriate for the particular model may be used within the scope of the present invention.

As discussed above, the training speech  
5 signals include additive noise that is partially removed by noise reduction model 408 to produce pseudo-clean data. One possible noise reduction technique that can be used under the present invention is spectral subtraction. In spectral  
10 subtraction, noise in the speech signal is sampled and the samples are converted to the frequency domain. The frequency content of the noise is then subtracted from a frequency representation of the speech signal to produce a pseudo-clean speech  
15 signal.

As shown in FIG. 5, the noise can be sampled from the speech data by sampling the speech signal during pauses in the actual speech. In FIG. 5, an example of a noisy speech signal is shown with  
20 time along horizontal axis 500 and the amplitude of the speech signal shown along vertical axis 502. In FIG. 5, the speech signal includes an active speech area 504 and two pauses 506 and 508. The active speech portion 504 of the speech signal has a higher energy content than the pauses 506 and 508. By  
25 sampling the speech signal during pauses 506 and 508, the background noise can be separated from the speech content of the signal.

FIG. 6 provides an example of the spectral  
30 content of noise samples taken during a pause in

0060950 10150  
009101 056950

speech such as pause 506 of FIG. 5. In FIG. 6, frequency is shown along horizontal axis 600 and the amplitude of each frequency component is shown along vertical axis 602. For noise spectrum 604 of FIG. 6, the spectral content has a higher magnitude in the middle band of frequencies and a lower magnitude at the lower and higher frequencies. During spectral subtraction, this frequency signature is used to generate a noise correction value for each frequency of the speech signal. The respective correction values are then subtracted from the corresponding frequency values of the speech signal to reduce the noise in the speech signal.

FIG. 7 provides a block diagram for one embodiment of noise reduction module 408 and feature extractor 410 of FIG. 4. In the embodiment of FIG. 7, noise reduction module 408 performs a spectral subtraction and feature extractor 410 produces Cepstral coefficients as its extracted features. In the embodiment of FIG. 7, noise reduction module 408 and feature extractor 410 are integrated together to form a single operating module. Although the functions of these two modules are integrated in FIG. 7, those skilled in the art will recognize that the individual components used to produce the embodiment of FIG. 7 need not be found on the same chip in hardware implementations of the invention or in the same software module in software implementations of the invention.





The correction values produced by weighting module 704 are stored in a memory 708 that is accessed by a spectral subtractor 710. Spectral subtractor 710 also receives the frequency domain values from FFT 700. For each frequency associated with the correction values stored in memory 708, spectral subtractor 710 subtracts the corresponding value in memory 708 from the frequency-domain value provided by FFT 700. This results in pseudo-clean frequency domain values at the output of spectral subtractor 710.

In other embodiments, the present invention  
25 uses Stereo Piecewise Linear Compensation for  
Environments (SPLICE) as the noise reduction  
technique. The SPLICE noise reduction technique is  
discussed in detail in a U.S. Patent Application  
entitled METHOD OF NOISE REDUCTION USING CORRECTION  
30 VECTORS, filed on even date herewith, having attorney

docket number M61.12-0325 and hereby incorporated by reference.

Under the SPLICE technique, noise is reduced by estimating the most likely clean feature vector that could be represented by a noisy feature vector from a noisy pattern signal. This is done by selecting a correction vector to add to the noisy feature vector to form the clean feature vector. To select the correction vector, the method determines which of a set of mixture components the noisy feature vector best matches. The correction vector associated with that mixture component is then added to the noisy feature vector.

Each correction vector is formed in part by subtracting a sequence of noisy channel feature vectors from a sequence of clean channel feature vectors, where the noisy channel and the clean channel contain the same speech signal, but where the noisy channel has additive noise. Typically, the correction vectors are formed before either the training data or test data are provided to the noise reduction module.

In one embodiment of the present invention, multiple sets of training data are used to incorporate multiple types of noisy environments in the training model. Thus, under one embodiment some training data is collected at a train platform, while other data is collected in a car, and still further data is collected in an airplane. FIG. 8 provides a flow diagram of one method under the present

In step 800 of FIG. 8, one set of training data with additive noise is created, by for example having a trainer speak in a selected noisy environment. In step 802, one or more noise reduction techniques are applied to the set of training data. The noise reduction techniques applied to the training data in step 802 can be the same for each type of noisy environment or may be tailored for the specific noisy environment upon which the noise reduction techniques are being applied.

If there are no other sets of data, the process continues at step 806 where the acoustic model is trained using all of the sets of pseudo-clean training data that result from the noise reduction techniques of step 802.

By using noise reduction techniques against multiple sets of training data that are each associated with different types of noise, the  
30 embodiments of the present invention produce more

FIG. 9 shows three probability distributions for a unit of speech. In FIG. 9, the feature vector from the speech signal is shown along horizontal axis 900 and the probability of the speech unit is shown along vertical axis 902. Probability distributions, 904, 906 and 908 of FIG. 9 are each associated with a different set of training data. Thus, each of the probability distributions is associated with a different type of additive noise. In FIG. 9, the sets of training data have not been passed through a noise reduction module.

Because probability distributions 904, 906 and 908 are spread out along the horizontal axis, the resulting combined probability distribution is flat. Such a distribution causes the model to be broadly defined so that the level of certainty in selecting the proper unit of speech is reduced since a given speech signal will produce similar probabilities for a large number of speech units.

FIG. 11 shows the effects of applying noise  
30 reduction to multiple sets of training data. In FIG.

11, the feature vector values after noise reduction are shown along horizontal axis 1100 and the probability of a unit of speech is shown along vertical axis 1102. In FIG. 11, the three probability distributions of FIG. 9 have been brought closer together by the noise reduction techniques. This results in distributions 1104, 1106 and 1108 respectively.

Because the individual distributions of FIG. 11 are brought closer together, the combined distribution 1200 shown in FIG. 12 is more sharply defined. Having such sharp definition in the probability distribution results in more certainty in the decision making process for selecting a unit of speech given an input speech signal. The sharpness of the definition is shown in distribution 1200 by the fact that the distribution rises quickly near a particular feature vector along the feature vectors of horizontal axis 1202 and provides a higher probability along vertical axis 1206.

In some embodiments where different noise reduction techniques are used for different sets of training data, the noise in the test data is sampled to determine which noise reduction techniques should be applied to the test data. FIG. 13 provides a block diagram of a noise reduction module 1300 for one such embodiment of the present invention.

In noise reduction module 1300, the noise in the input test speech signal is sampled by a noise sampler 1301, by for example using the technique

0066950-101600

5

10

15

25

30 shown in parallel in FIG.13, in other embodiments,



5 to apply to each segment of test data.

Although the present invention has been  
20 described with reference to particular embodiments,  
workers skilled in the art will recognize that  
changes may be made in form and detail without  
departing from the spirit and scope of the invention.

Although the present invention has been  
20 described with reference to particular embodiments,  
workers skilled in the art will recognize that  
changes may be made in form and detail without  
departing from the spirit and scope of the invention.